GPT4Scene: Understand 3D Scenes from Videos with Vision-Language Models

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, Hengshuang Zhao

香港大學 THE UNIVERSITY OF HONG KONG

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

ICLR International Conference On Learning Representations

Pilot Study

Prompts of Closed-sourse VLMs.

System Prompt

You are a 3D indoor scene assistant. We provide **a labeled 2D image** and **a labeled Bird's Eye View (BEV) image** for analysis.
1. The 2D image has 8 frames captured at equal intervals from a video, arranged in a **2x4 grid** from left to right, top to bottom.
2. Object labels are numbered, with **numbers matching between the 2D and BEV images to indicate the same objects**.

ScanQA Prompt

You are now required to provide answers based on the given questions.

Important Guidelines

1. **When answering questions, do not reference the marks directly.** These marks are only provided to assist in understanding the layout. Your answers should refer to specific objects in the scene, not the marks.
2. When describing directions or positions, use prominent objects in the image to express spatial relationships, and do not refer to labels.
3. **Keep your answers as concise as possible.** For questions regarding color, quantity, etc., aim for **1-5 words**. For questions about spatial relationships, answers can be slightly longer but should not exceed 10 words. Do not provide any additional, irrelevant information.

Answer Format

1. All answers must be in lowercase. Answers should not include any punctuation marks. Any numbers mentioned must be in Arabic numerals.
2. **Please format your answers as follows: '##Q1## answer1, ##Q2## answer2, ...'.**

Examples:
- Question: What color table is on the left side of the cabinet?
- Answer: light brown
- Question: What is on the left of the tv?
- Answer: bicycle on floor

Zero-shot Prompting

1. Syetem Message: <System Prompt> + <ScanQA Prompt>
2. User Message (image type): <url_for_frames> + <url_for_BEV>

3. Query Message: 'What is the black chair in front of?'

Refinement Procedures

1. Get the answer: '##Q1## White board.'
2. Remove answer format: 'White board.'
2. Refinement and clean the answer:
· Remove singular and plural forms.
· Remove unnecessary adjectives.
· Remove punctuation and spaces.
· Remove uppercase and lowercase distinctions.

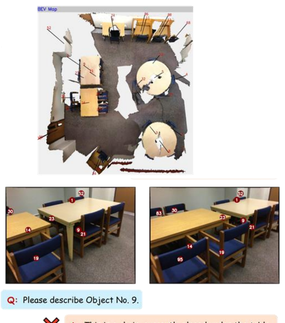
Final Results: 'whiteboard'

Prompts of Closed-sourse VLMs. We show the prompts used for GPT-4o (GPT4Scene), which consist of a system prompt and a benchmark prompt. After generating responses, we further refine them.

Failure Cases

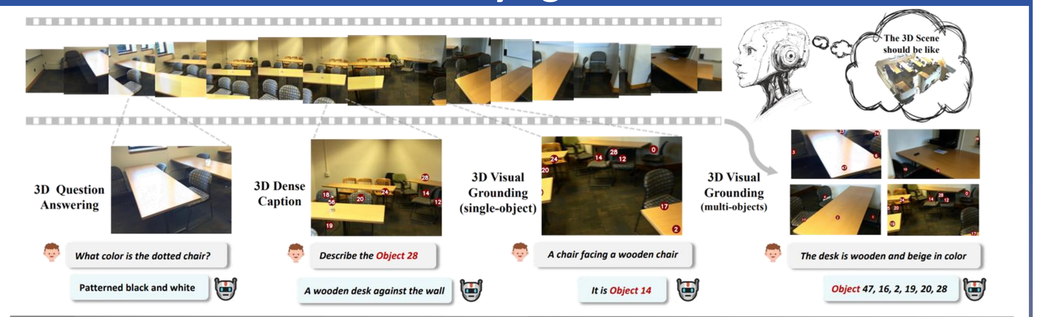
Table 1: **The zero-shot capability of GPT4Scene.** Video + GPT4Scene Inference without Fine-tuning.

Zero-shot 3D QA	ROUGE@ScanQA		EM-1@SQA3D	
	VID	+4Scene	VID	+4Scene
3D LLM Based Model	*Pre SOTA*		*Pre SOTA*	
Chat-scene Huang et al. (2024a)	*41.6*		*54.6*	
Open-sourced VLM Based Model				
Qwen2-VL-2B Wang et al. (2025b)	29.5	30.0 +0.5	36.9	36.2 -0.7
Qwen2-VL-7B Wang et al. (2025b)	30.8	33.2 +2.4	42.1	43.1 +1.0
Qwen2-VL-72B Wang et al. (2025b)	32.1	35.1 +3.0	41.5	44.0 +2.5
Closed-sourced VLM Based Model				
GPT-4o OpenAI (2024)	34.2	39.3 +5.1	42.0	44.8 +2.8
Gemini-1.5-Pro Team (2024a)	35.1	39.4 +4.3	43.5	46.0 +2.5



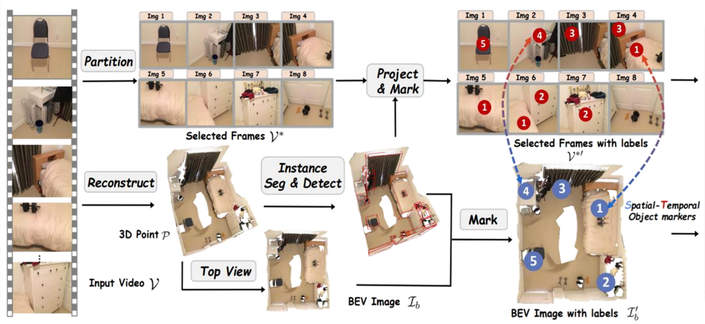
Closed-source VLMs still show limited zero-shot 3D understanding, with relatively low scores and frequent failures in object correspondence.

Motivation: Understand 3D by ego-centric video



GPT4Scene understands 3D scenes and performs tasks like 3D question answering, dense captioning, and visual grounding **using only video input**. Unlike 3D point LLMs, it relies solely on vision, with global context provided by a BEV image.

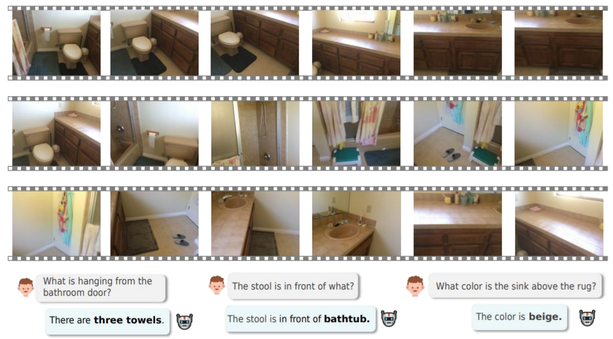
Using Markers to Maintain Correspondence



Training pipeline:

- Input video sequences
- Key frame selection
- 3D scene reconstruction
- Instance segmentation
- Top-view BEV generation
- Object projection to frames
- Cross-view marker assignment
- Spatial-temporal object markers
- Frame–BEV correspondence
- Structured input to VLM

Experiments Results



Qualitative Results GPT4Scene answers object-centric 3D questions and attends to the correct spatial regions across frames. Below is attention map.

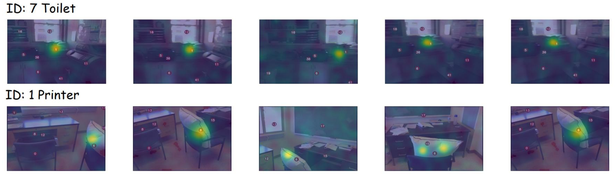


Table 4: **Evaluation of 3D dense caption on Scan2Cap Chen et al. (2021).** Our results outperform those of existing 3D LLM based models.

3D Dense Caption	IoU@0.25		IoU@0.5	
Methods	BLEU-4	ROUGE	BLEU-4	ROUGE
Task-Specific Model				
Scan2Cap Chen et al. (2021)	34.2	55.3	23.3	44.5
3DJCG Cai et al. (2022)	40.2	59.2	31.0	50.8
X-Trans2Cap Yuan et al. (2022)	35.7	54.7	25.1	45.3
3D-VisTA Zhu et al. (2023)	36.5	57.6	34.0	54.3
Vote2Cap-DETR Chen et al. (2023b)	39.3	59.3	34.5	54.4
3D LLM Based Model				
LL3DA Chen et al. (2024b)	41.4	59.5	36.8	55.1
PQ3D Zhu et al. (2024b)	–	–	36.0	-
LEO Huang et al. (2024c)	–	–	36.9	57.8
Chat-scene Huang et al. (2024a)	38.2	60.6	36.3	58.1
Grounded 3D-LLM Chen et al. (2025)	–	–	35.5	-
Robin3D Kang et al. (2025a)	–	–	38.4	–
Ross3D Wang et al. (2025a)	–	–	43.4	66.9
Vision LLM Based Model				
LLaVA-3D Zhu et al. (2025)	–	–	41.1	63.4
Video-3D-LLM Zheng et al. (2025)	–	–	41.3	-
InternVL3-8B (GPT4Scene)	44.1	63.1	41.4	60.3
Qwen2-VL-7B (GPT4Scene)	43.1	61.9	40.6	59.3
Qwen2.5-VL-7B (GPT4Scene)	**45.9**	**67.9**	**44.1**	**67.1**

Table 5: **Evaluation of 3D visual grounding on ScanRefer Chen et al. (2020) and Multi3DRef Zhang et al. (2023c).** Our method reaches state-of-the-art performance over all methods for the 3D visual grounding task.

3D Visual Grounding	ScanRefer		Multi3DRef	
Methods	Acc@0.25	Acc@0.50	all F1@0.25	all F1@0.50
Task-Specific Model				
3DVG-Transformer Zhao et al. (2021)	47.6	34.7	–	25.5
3DJCG Cai et al. (2022)	49.6	37.3	–	26.6
D3Net Chen et al. (2022a)	–	37.9	–	32.2
M3DRef-CLIP Zhang et al. (2023c)	51.9	44.7	42.8	38.4
3D LLM Based Model				
3D-LLM Hong et al. (2023)	30.3	-	-	-
Grounded 3D-LLM Chen et al. (2025)	47.9	44.1	45.2	40.6
Chat-scene Huang et al. (2024a)	55.5	50.2	57.1	52.4
Ross3D Wang et al. (2025a)	61.1	54.4	59.6	54.3
Vision LLM Based Model				
LLaVA-3D Zhu et al. (2025)	54.1	42.4	-	-
Video-3D-LLM Zheng et al. (2025)	58.1	51.7	58.0	52.7
InternVL3-8B (GPT4Scene)	63.4	57.7	65.5	60.7
Qwen2-VL-7B (GPT4Scene)	62.6	57.0	64.5	59.8
Qwen2.5-VL-7B (GPT4Scene)	**65.6**	**59.5**	**67.3**	**62.8**